Document dissimilarity within and across languages: A benchmarking study

Richard S. Forsyth and Serge Sharoff
University of Leeds, UK

Abstract

Quantifying the similarity or dissimilarity between documents is an important task in authorship attribution, information retrieval, plagiarism detection, text mining, and many other areas of linguistic computing. Numerous similarity indices have been devised and used, but relatively little attention has been paid to calibrating such indices against externally imposed standards, mainly because of the difficulty of establishing agreed reference levels of inter-text similarity. The present article introduces a multi-register corpus gathered for this purpose, in which each text has been located in a similarity space based on ratings by human readers. This provides a resource for testing similarity measures derived from computational text-processing against reference levels derived from human judgement, i.e. external to the texts themselves. We describe the results of a benchmarking study in five different languages in which some widely used measures perform comparatively poorly. In particular, several alternative correlational measures (Pearson *r*, Spearman rho, tetrachoric correlation) consistently outperform cosine similarity on our data. A method of using what we call 'anchor texts' to extend this method from monolingual inter-text similarity-scoring to inter-text similarity-scoring across languages is also proposed and tested.

Correspondence:
Richard S. Forsyth,
Centre for Translation Studies, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT, UK.
E-mail:
forsyth_rich@yahoo.co.uk

1 Introduction

Quantifying the similarity or dissimilarity between documents is a problem that arises in authorship attribution (Juola, 2006), corpus comparison (Kilgarriff, 2001), information retrieval (Salton and McGill, 1983), near-duplicate detection (Chowdhury *et al.*, 2002), plagiarism detection (Clough and Gaizauskas, 2009), term extraction (Li and Gaussier, 2010), text mining (Weiss *et al.*, 2005), and many other natural-language processing tasks.

Many indices have been proposed and used for such purposes, but comparatively little effort has been devoted to calibrating such indices in the sense of systematically comparing the outputs of various textual (dis-)similarity functions with some kind of text-external standard. A likely reason for this is that, although texts can be, and have been, placed into categories on the basis of genre, register, topic, and other discourse-level attributes, there is no widespread agreement on how similar, or otherwise, these different categories are (Wu *et al.*, 2010). For example, Category J in the Brown Corpus (Kucera and Francis, 1967) contains research articles in radio engineering, chemistry, and psychoanalysis, as well as essays on opera and poetry. Arguably there is more dissimilarity among texts within this category than between samples drawn from two separate categories L (mystery and crime fiction) and N (adventure fiction).

Literary and Linguistic Computing, Vol. 29, No. 1, 2014.
doi:10.1093/llc/fqt002 Advance Access published on 6 February 2013

The present article addresses this issue by introducing the *Pentaglossal Corpus,* a collection of texts in five languages, where each text has a location in a two-dimensional similarity space derived from the ratings of human readers. Thus each document can be located on the basis of readers' assessment of its contents. This then provides the grounding for a benchmarking study that compares the dissimilarities derived from this reader-generated text-external framework with several text-internal measures of dissimilarity. The main aim of this investigation is to find a robust inter-text dissimilarity-scoring function, and to generalize it from the monolingual to the multilingual case.

1.1 Related work

Work related to the present investigation has been carried out in a number of subfields. Kilgarriff (2001) has studied methods of assessing comparability between corpora, but only in a monolingual context and at the corpus level. Juola (2006) has found that similarity-based methods perform well in authorship attribution trials, but again only described tests within, rather than across, languages. Both Li and Gaussier (2010) and Su and Babych (2012) have tested techniques that quantify corpus and textual dissimilarities across languages. Their experiments differ from the present study in requiring bilingual dictionaries for the languages concerned. In the field of cross-language information retrieval, Chen and Bau (2009) describe retrieval mechanisms that find documents semantically related to a query text. These procedures perform inter-text similarity ranking across languages, but they rely on Google's proprietary translation algorithms. Potthast *et al.* (2011) survey a range of approaches to cross-language plagiarism detection, which implicitly or explicitly compute inter-text similarities. They compare three particular techniques empirically. However, the method (the simplest) that they found to give best results, based on character trigrams, would be problematic to extend to texts in other than the Roman alphabet. Banchs and Costa-Jussà (2010) propose a method of cross-language sentence-matching (which could be extended to document-matching) and test it on sentences from the Spanish Constitution in the original Spanish and in translations into four other languages, including Basque and English. Like ours, their approach uses a collection of 'anchor documents'. It differs by interposing a stage that requires the computation of explicit semantic maps from the anchor documents of each language.

A significant aspect in which the present study differs from most related work, including those cited previously, is in establishing an external reference criterion of inter-text dissimilarity that can be regarded as an interval scale, rather than simply relying on category membership for evaluation purposes.

1.2 Outline

In Section 2 of this article, we describe our main test data set, a corpus in five languages. In Section 3, we describe a scheme based on readers' judgements that enables the setting up of a similarity space in which documents from this corpus can be located, and thus the derivation of an external criterion of inter-text dissimilarity. We also present some evidence relating to the reliability of this external criterion. Section 4 outlines the textual features used in our experiments. Section 5 describes an experiment aimed at discovering which of a number of plausible inter-text dissimilarity indices correlate best with the external criterion. In Section 6, we show how text similarity can be estimated indirectly, by using 'anchor texts'. This allows us to move, in Section 7, from monolingual similarity-scoring to cross-language similarity-scoring. Finally, the Discussion briefly considers the implications of this research and outlines some future directions.

2 A Pentaglossal Corpus

For this benchmarking exercise, we have assembled a parallel corpus comprising 113 texts in five languages, namely, English, French, German, Russian, and Chinese—the *Pentaglossal Corpus*. Each text has a translation equivalent in the other four languages, which allows us to calibrate our dissimilarity measures. Li and Gaussier (2010) also tune their multilingual comparability procedure by using Europarl, a parallel corpus of European Parliament

Table 1 Pentaglossal Corpus composition (April 2012)

Type	Documents	Tokens	Description
Bib1	5	5,503	Bible, Old Testament extracts
Bib2	6	10,140	Bible, New testament extracts
Corp	6	5,074	Corporate statements of self-promotion
Fict	30	138,704	Fiction: novel chapters or short stories
Marx	5	31,499	Marxist documents
News	10	7,078	News articles
Opac	3	3,766	Open access declarations
Tedi	11	22,758	Transcripts from Ted.com initiative
Tele	14	44,856	Telematics, engineering
Teli	1	2,733	Telematics, instructions
Tels	15	8,974	Telematics, software
Unit	4	19,205	United Nations documents
Wind	3	7,417	Wind energy articles

proceedings (Koehn, 2005). However, in comparison with the World Wide Web or any other diverse text collection (like the BNC or Brown Corpus), Europarl is homogeneous in terms of its topics and genres, so it is difficult to generalize any results obtained from it. Thus the Pentaglossal Corpus contains texts from a mixture of domains and registers, as outlined in Table 1.

The texts in the Pentaglossal Corpus are documents, or excerpts from documents, that have been classified into thirteen text types, as listed in the table. The coding scheme can be collapsed to ten classes by ignoring the fourth character of the four-character codes. Table 1 summarizes some basic attributes of this corpus, including number of texts and number of word tokens. The word count shown is from the English version, which contains 3,07,707 word tokens altogether, according to our tokenizer.

To the best of our knowledge, these texts are out of copyright or covered by Creative Commons or similar licences, so the corpus can be made available to other researchers. The release of 4 April 2012 can be found at the following address:

http://corpus.leeds.ac.uk/tools/5gcorpus.zip

The main point about this corpus is that it allows us to impose an external reference level of similarity between documents. Each text is associated with metadata concerning its provenance, as well as with two coordinates (horz, vert) that give the location in a two-dimensional similarity space. This provides information from which a criterion level of dissimilarity can be computed for any pair of documents. This is illustrated in Fig. 1, in which each text is represented by its category label.

3 Establishing Reference Levels of Document Dissimilarity

The procedure used to arrive at these locations was as follows. Three volunteers read all 113 documents, in English. Each document was then rated on seventeen textual attributes, using a four-point scale: 0 meaning the attribute was absent, 0.5 meaning it was present but only to a small extent, 1 meaning it was somewhat or partly present, and 2 meaning the text was strongly characterized by the attribute in question. The attributes and their descriptions are summarized in the Appendix. As an index of inter-rater reliability, we used Krippendorff alpha, in preference to Cohen or Fleiss kappa, because alpha is more general (Kripendorff, 2004), because it handles more than two judges naturally, and, in particular, because it takes account of the magnitude of differences between judges, not just the fact of agreement or disagreement. Using the ReCal web service of Deen Freelon (Freelon, 2010), on an interval scale, Krippendorff alpha was computed as 0.764 between the three judges, which we regard as an acceptable level of reliability.

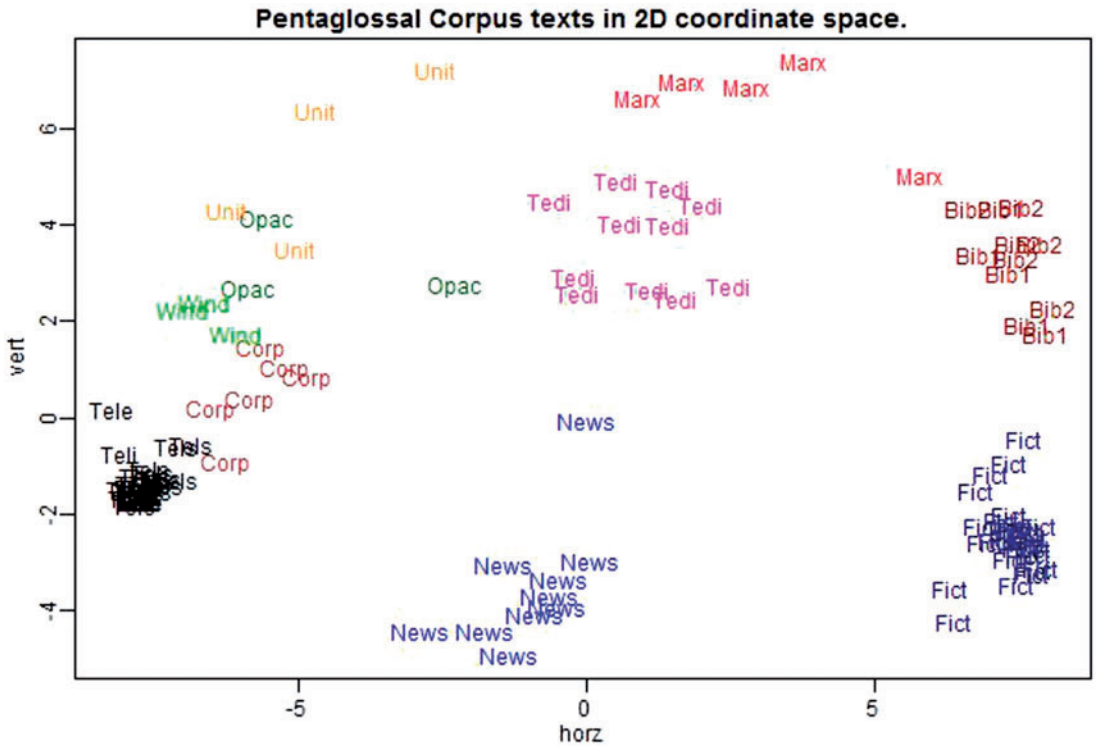

Fig. 1 Pentaglossal Corpus texts in two-dimensional similarity space

The scores of the three judges were converted into a 113-by-17 grid by simple summation. This was reduced to a 113-by-16 grid by dropping the column/attribute on which the three judges agreed least (number 6, informal language). From this, a 113-by-113 distance matrix was produced and then processed by the Sammon multidimensional scaling procedure (Sammon, 1969) in the Modern Applied Statistics with S-Plus (MASS) library of the R package (R Development Core Team, 2009). This gave a two-dimensional solution with a stress level of 0.0264, indicating, in effect, that >97% of the information in the distance matrix is preserved in the x and y coordinates. These coordinates were multiplied by ten and rounded to two decimal places and then exported to the metadata file as (horz, vert) coordinates.

These coordinates are derived from human judgement at the level of conceptual content, and thus can be compared with results from low-level text-processing, which is the aim of the present exercise. Moreover, they provide the basis for a criterion of inter-document difference that is not merely binary (i.e. not just same-versus-different category).

Figure 1 shows the positions of all 113 documents in this conceptual similarity space. Each text is represented by the first four characters of its name, which indicates its text-type. An inspection of Fig. 1 reveals that the results accord well with more general intuitions about the closeness or difference of the various text types. It would seem uncontroversial that the three types of text dealing with telematics should fall near to each other, that the Old and New Testament biblical extracts should be relatively close, that the telematic documents should be far distant from the fictional extracts, and so on. It is tempting to assign semantic labels to the axes of this graph, although we resist that temptation at present, pending further investigation of the contents of these and other texts.

Note that, although texts in the same category do tend to gravitate together, some text types are less tightly clustered than others, and that some of our classes overlap, thus confirming the importance of giving each text its own individual location.

Essentially Fig. 1 is a visual representation of our target. What we seek is a low-level text-processing procedure that matches this configuration as closely as possible.

As each English text has a translation equivalent in the other four languages, we give the same horizontal and vertical coordinates as in English to the texts in Chinese, French, German, and Russian. It is reasonable to presume that a good translation should preserve most of the relevant properties of an original text; however, some differences are to be expected (Banchs and Costa-Jussà, 2010). To gain an idea of whether differences between languages were serious enough to undermine the justification for assigning the same coordinates to all five languages, we conducted a further calibration exercise. Two further judges, speakers of Chinese and Russian, were given a random subset of seventeen texts from the Chinese and Russian Pentaglossal Corpus, respectively, to rate on the same attributes as used in rating the English texts (see Appendix). Then Krippendorff alpha (Kripendorff, 2004) was calculated for the 289 ratings made on these seventeen documents among three judges, two individuals and one composite: the Chinese judge, the Russian judge, and the arithmetic mean score from the three English judges. The value of alpha (interval-scale data) thus computed was 0.733. As this is close to the value obtained when comparing the three English judges (0.764), our working hypothesis is that it unlikely that serious imprecision is introduced by assigning scores derived from the English consensus to the other four languages.

4 Feature-Finding for Document Dissimilarity

Having constructed a target to aim at, we can examine various ways of computing textual similarity and test how well they match the target. For this purpose, we used the 113 text files of each language without any linguistic pre-processing, except tokenization.

The features used in this study to characterize texts pay more respect to the inescapably sequential nature of language than the more conventional term-vector (or 'bag-of-words') approach. This is an attempt to exploit what Sinclair (1991) calls the 'idiom principle', namely, the tendency for speakers and writers, as well as listeners and readers, to work with chunks of language rather than isolated words. The results of such chunking have been referred to by a variety of terms, such as 'collocations', 'concgrams', 'flexigrams', 'lexical bundles', 'multi-word expressions', 'prefabricated phrases', 'skipgrams', among other designations (Biber *et al.*, 2004; Cheng *et al.*, 2006; Min and McCarthy, 2010). All are generalizations of the basic notion of an n-gram, but different authors have generalized this concept in slightly different ways, and thus the meanings of these terms overlap in a somewhat confusing manner. As the terminology for flexible multi-element linguistic units is not yet standardized, we refer in this article to 'elastigrams'.

A program in Python3 has been written to generate a list of elastigrams from a given collection of text files. A short extract from its output on the French part of the Pentaglossal Corpus is shown in the following text. This illustrates the kind of linguistic fragments extracted by the algorithm.

('de', 'que', 'le')	5 40 0.00310
('ne', 'pas', 'que')	5 38 0.00304
('par', 'le', 'de')	5 47 0.00303
('sur', 'la', 'du')	5 37 0.00293
('de', 'la', 'que')	5 43 0.00290
('la', 'première', 'fois')	5 38 0.00289
('que', 'je', 'ne')	5 41 0.00289
('dans', 'le', 'monde')	5 48 0.00286
('ce', 'que', 'vous')	5 43 0.00283
('le', 'de', 'sa')	5 39 0.00281
(''c'est'', 'à', 'dire')	5 54 0.00278
('pour', 'plus', ''d'informations'')	5 53 0.00277
('de', 'la', 'vie')	5 46 0.00274

Here the first numeric column gives the window width (5), i.e. number of tokens within which the three specified tokens must be found. The next column is the raw frequency of the elastigram within the whole corpus. The last column is the value on which the items are sorted. This is a

ubiquity measure (u) based on the adjusted frequency proposed by Rosengren (1971) but modified for unequal block size:

$$u = (\sum_{j=1}^{n} w_j \times \sqrt{r_j})^2$$

where w_j is the square root of the length of document j divided by the total of the square roots of all document lengths, and r_j is the percentage rate of occurrence of the elastigram in that document.

In our terms, the items listed above are 3/5-grams. They can be thought of as simple patterns that match three tokens within a window of five tokens. These tokens need not be consecutive, although in the ordered case they must be sequential. With a 3/5-gram, up to two other tokens can intervene between the three specified tokens, and the pattern will still match. (The software can deal with unordered elastigrams, where the order does not matter, but these are not used in the present article.) Note that a 1/1-gram is a single token, thus this framework does, of course, allow more traditional word-based analyses.

For n/m-grams where n and m are more than small numbers, the number of elastigrams in a corpus can become huge. Most of these occur only once or twice. To deal with this combinatorial explosion, the elastigram-finding program generates all elastigrams of the requested size in each text, but only keeps the most frequently occurring K in that text, where K is the rounded square root of the text length in tokens. The union of these sets from all documents in the corpus is sorted in descending order of adjusted relative frequency in the corpus as a whole, i.e. the ubiquity measure defined previously (square of the weighted mean root percentage occurrence rate).

The top N_f elastigrams are retained for output as defined in

$$N_f = int(\sum_i \ln(t_i) + 0.5)$$

where t_i is the size of document i in tokens. These N_f items form a vocabulary that will be used as the feature-set for subsequent processing. (The actual numbers of elastigrams retained in the present experiment for each language were: de = 836, en = 839, fr = 845, ru = 819, zh = 886.) To compute inter-text dissimilarity, each document is represented as a numeric vector, where the numbers are frequencies of occurrence of each elastigram, or values derived from those frequencies.

5 An Experiment on Monolingual Document Dissimilarity

A major goal of this experiment was to examine the quality of a variety of text-based dissimilarity measures, i.e. how well they match the text-external dissimilarity criterion defined in Section 3. As our response variable, we used the product-moment correlation (r) of the dissimilarity derived from readers' judgements, with the dissimilarity computed by a variety of distance functions (distmode) applied to a variety of transformations of the raw frequency data (varmode) – on the five languages of our corpus. Table 2 summarizes the predictive factors investigated in our experiment.

It should be noted that Kullback–Leibler divergence is only valid for probabilities, and therefore should strictly speaking only be used on rates (i.e. with varmode = ra). Moreover, it is undefined if any probability in the second vector is zero. Thus the function used here is actually a modified form of K-L divergence, amended as follows: half the smallest non-zero value in the data vector is added to all values in that vector, and then the augmented values are divided by their sum total to create pseudo-probabilities. As this is an inherently asymmetric measure, to ensure symmetry, we always used the vector derived from the larger text as the reference distribution (q).

It should also be noted that not all the 72 combinations of distmode and varmode give distinct results. For example, rr+icor must give the same value as fr+irho and ra+irho. However, for simplicity, each possible combination was tested. For each of the five languages, all combinations of the predictive factors listed previously were used to generate dissimilarity scores for each distinct pair of texts (6,328 scores per combination), which were then correlated with the 6,328 reader-derived criterion dissimilarities to yield a quality score. This

Table 2 List of experimental factors

Factor	Levels
Language	DE, EN, FR, RU, ZH
Elastigram size	1/1, 2/3, 3/5
Contents of data vector (varmode)	fr: raw frequencies of each elastigram in the feature-list. bi: binarized frequencies, i.e. one if the elastigram concerned is present, otherwise zero. ra: percentage occurrence rates of each elastigram, i.e. relative frequencies. rf: square roots of raw frequencies. ri: riditized frequencies (Bross, 1958). rr: reciprocal of rank position of each elastigram by relative frequency. rx: $1 - 1/(1 + \sqrt{f_j})$ where f_j is frequency of elastigram j. wt: tf-idf weights (Weiss *et al.*, 2005).
Distance function (distmode)	canb: Canberra metric $d(p,q) = \sum_{j=1}^{n} \frac{(\lvert p_j\rvert - \lvert q_j\rvert)}{(\lvert p_j\rvert + \lvert q_j\rvert)}$ (with zero result when denominator equals zero). city: city-block distance (Minkowski L_1-metric). czed: Czekanowksi coefficient (Everitt, 1998) $d(p,q) = 1 - 2 \times \sum_{j=1}^{n} \min(p_j,q_j) + \sum_{j=1}^{n}(p_j \div q_j)$ eudi: Euclidean distance. icor: inverse product-moment correlation coefficient (1-r). icos: inverse cosine similarity (1-cos). irho: inverse of Spearman's rank correlation coefficient (1-rho). itet: inverse tetrachoric correlation coefficient (1-tc) estimated according to Karl Pearson's formula (Upton and Cook, 2008) $d(p,q) = 1 - \sin(\frac{\pi}{2} \times (\sqrt{ad} - \sqrt{bc}) \div (\sqrt{ad} + \sqrt{bc}))$ where a, b, c, d are counts in a fourfold table constructed by reference to the median values in the vectors such that a is the number of times both values exceed their median, b is the number of time the first value exceeds its median while the second does not, c is the number of times the second value exceeds its median while the first does not, and d is the number of times neither value exceeds its median. (In fact, all four counts were incremented by 1 as an attenuation factor to avoid zero cell counts). kuld: Kullback–Leibler directed divergence (Kapur and Kesavan, 1992) $d(p,q) = \sum_{j=1}^{n}(p_j \log(p_j/q_j))$.

produced 1,080 quality scores in total (five languages, three elastigram sizes, eight variable modes, and nine distance functions).

To disentangle the effects of the four predictor variables, we used the recursive partitioning function in the conditional inference tree package, party (Hothorn *et al.*, 2006), from the R library, to grow a regression tree with corscore (Pearson's *r*) as the response variable. Maximum depth was set to three. This created the symmetric tree structure shown in Fig. 2 with eight leaf nodes. The partitioning algorithm uses distance mode as its primary split, suggesting that the distance function is the most important factor, with elastigram size (subsize being the first number in an n/m-gram) as the splitting factor at the next level down.

To interpret such a tree, note that the oval nodes identify features, whereas the lines connecting nodes signify tests made on those features. Hence, for example, following the left-hand branches of this tree from top to bottom, we reach the subset of cases in which distmode is icor, irho, or itet; subsize is less than or equal to two; and distmode is irho or itet (thus filtering out icor at the lowest level). The rectangles at the foot of this tree display box plots of the corscore values found at the leaf nodes of the tree, i.e. for each particular combination of truth-values resulting from the tests performed on the branches leading to that leaf node. In each box plot, the median value of the response variable is shown as a dark horizontal line. In this diagram, the better results appear toward the left-hand side. In a nutshell, this tree indicates that the only distance modes worth considering seriously for the present task are icor, irho, and itet; and that 3/5-grams should be avoided, except perhaps with Chinese (nodes 4, 5, and 7).

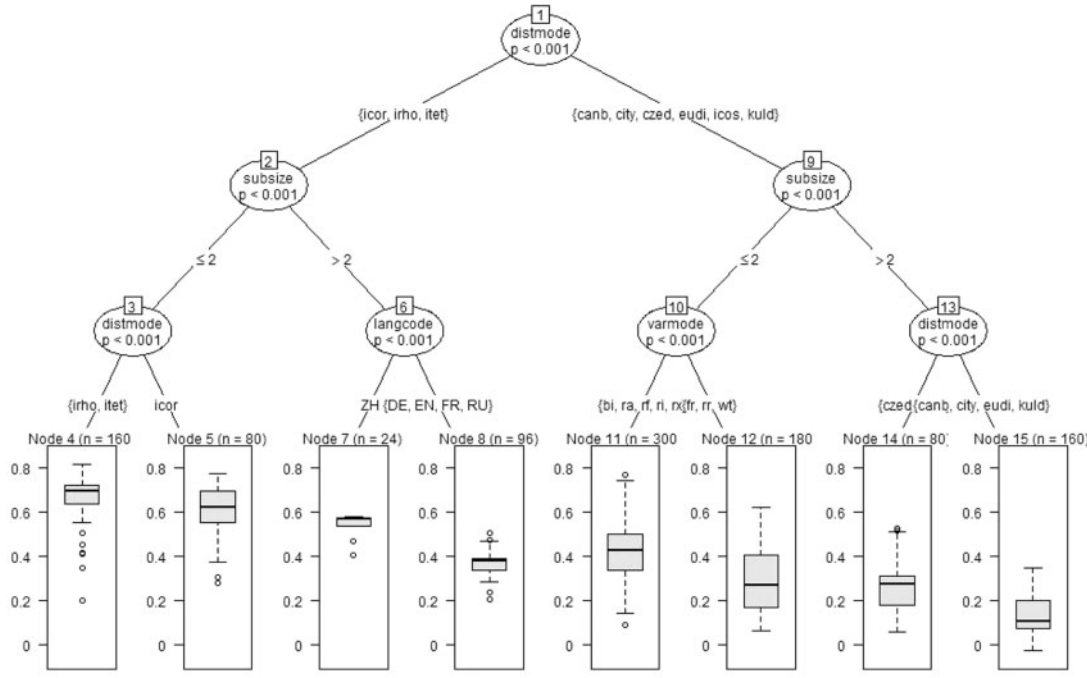

Fig. 2 Regression tree with corscore as response variable showing that choice of distance function (distmode) is the most important factor influencing correlation between humanly assigned and textually computed levels of dissimilarity (corscore)—with length of elastigram (subsize) as the next most important factor

The fact that distance mode is chosen as the root decision-variable implies that it is the most important of our four factors in determining the quality of dissimilarity-scoring. The simple distance measures (Canberra, City-block, Czekanowski, and Euclidean distance) would seem to be unsuitable for the present task. City-block and Euclidean distance were not expected to perform well on linguistic data, which are characterized by skewed distributions with high variance, but the presence of the more sophisticated cosine score (icos) among the 'also-rans' is somewhat unexpected.

K-L divergence also performs relatively poorly with these data. Arguably, it is unfairly penalized by being given inappropriate inputs in most conditions, although, against this, the best result for K-L divergence was not with the variable mode for which it is designed (ra) but with simple binarization (bi). It is likely that K-L divergence could only be used effectively in this context with a complicated Bayesian pre-processing phase designed to give accurate probability estimates from skewed small-sample frequency vectors.

As the longest elastigram size (3/5) would appear to be contraindicated, Fig. 3, which shows the interaction of distmode and varmode data for 1/1-grams and 2/3-grams only, excluding 3/5-grams, can be used to give a visual impression of how the two most influential variables interact in determining correlational quality.

This illustrates a tangle of interaction effects: although there is no 'star performing' data transformation (varmode), certain variable modes are suited by, or unsuited to, certain distance functions (distmode). Something that is clear from this diagram is that the classical distance measures lead to worse performance than the (inverse) correlational measures, with modified K-L divergence occupying an intermediate position. Perhaps surprisingly, binarized elastigram frequencies (simple presence/absence data) perform relatively well, especially in combination with inverse correlation or

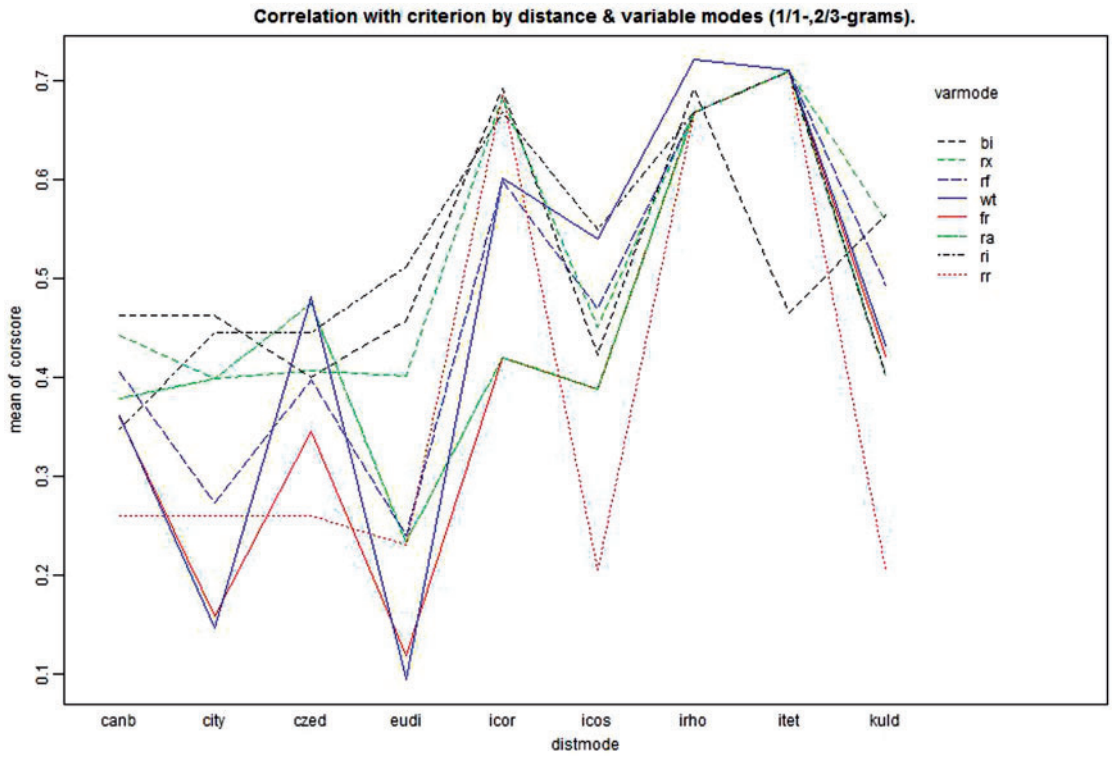

Fig. 3 Interaction of distance mode and data-vector mode in determining quality score

inverse rank correlation, although not with inverse tetrachoric correlation (itet). For itet, it is obviously better to binarize at the median, as was done in this experiment.

Another perspective on these results can be gained by considering the ninety-eight cases (of 1,080) where the response variable was 0.7071 or higher, i.e. a respectable performance accounting for at least half the variance in the target variable.

The figures in the last column of Table 3 point toward an intriguing linguistic effect. On the whole, the matching between text-external and text-internal dissimilarity was best in Chinese and worst in Russian, with the other languages intermediate. Thus this approach does least well with the most inflected of these five languages and best with the least inflected. This suggests that lemmatization might be helpful when dealing with highly inflected languages, or, better still (if feasible), some kind of morphological decomposition. From the point of view of recommendation, it would seem reasonable to recommend avoidance of long elastigrams (3/5-grams) and, having done that, to pick one of the three best combinations of variable mode and distance mode, namely, wt+irho, fr+itet, and bi+icor. In fact, with itet, any variable mode other

Table 3 Factor values associated with best 98 cases (of 1,080)

Distance function (distmode)	Data transformation (varmode)	Gram size	Language
itet 49	ri 15	1/1 62	zh 41
irho 30	rr 14	2/3 36	fr 20
icor 17	rx 14	3/5 0	de 19
eudi 1	wt 14		en 18
icos 1	rf 11		ru 0
(others) 0	bi 10		
	fr 10		
	ra 10		

than bi gives equivalent results, but fr involves the simplest computation.

A straight comparison of the two leading combinations wt+irho and fr+itet in equivalent conditions regarding language and span size throws some light on the worth of the tf-idf transformation. Of ten comparable cases, wt+irho yielded a higher corscore in eight, with a mean improvement of 0.0117. However, this difference was not statistically significant (paired *t*-test: $t = 1.79$, $P = 0.107$). Hence the improvement bought by using the tf-idf transformation, which requires information from an entire corpus, compared with simply using frequencies, which require information only from the two texts under consideration, would appear to be scarcely worthwhile. In this connection, it is quite striking that the cosine measure, which, along with tf-idf term weighting, has been a standard in information retrieval for several decades (Sparck Jones, 1972; Salton and McGill, 1983), is clearly suboptimal. Indeed, in every variable mode, including the mode for which it was designed (wt), icos was outperformed by icor, irho, and itet.

6 Indirect Inter-Text Similarities, Using Anchor Texts

Our purpose in establishing a robust measure of monolingual inter-text dissimilarity is to use it as a stepping-stone to quantifying similarity or difference between texts in different languages. Most attempts to do this make use of bilingual or multilingual lexicons or thesauruses (e.g. Steinberger *et al.*, 2002; Chen and Bau, 2009; Su and Babych, 2012), but it can be achieved by other means. To do so, we develop an idea suggested by Rajman and Hartley (2001) in the context of assessing translation quality. In this approach, the similarity of one text to another is not computed directly but is estimated from their profiles of distances (or similarities) to a collection of other texts, which we refer to as 'anchor texts'.

The underlying assumption is that if two texts are similar to each other, and are compared with a set of other documents, the 'anchor texts', then they will have a similar pattern of similarity scores to those anchor documents. If the two texts being indirectly compared are dissimilar, then the profiles of their similarity scores to the anchors will differ. A small numerical example is illustrated in Table 4.

Table 4 Small-scale anchor-distance matrix

Anchors	A1	A2	A3	A4	A5
X	0.75	0.22	−0.48	0.00	0.96
Y	0.25	0.22	0.89	0.17	0.04
Z	0.69	0.42	−0.40	0.04	0.95

Here we consider just five anchors and three domain texts X, Y, and Z. Each row contains the similarities between the domain text and the anchors. (Dissimilarities would give the same result.) The correlation (Pearson *r*) between row X and row Y is −0.72, to two decimal places. The correlation between rows X and Z is 0.99, to two decimal places. Thus we assume that text X and Y are dissimilar, whereas X and Z are highly similar.

To test how well this line of reasoning is likely to work out in practice, before introducing the complicating factor of different languages, we conducted a small-scale monolingual trial using the Brown and Lancaster Oslo Bergen (LOB) corpora (Kucera and Francis, 1967; Hofland and Johansson, 1982). Both these corpora consist of 500 texts of approximately 2,000 words each categorized under fifteen subject headings. The Brown corpus samples written American English of the early 1960s, whereas the LOB corpus uses, essentially, the same category scheme to sample British English of the early 1980s.

In our first monolingual test, we used LOB as the domain corpus ($n = 500$), with the Pentaglossal Corpus as the anchors ($n = 113$). First, a matrix of inter-text dissimilarities was computed on the LOB corpus using the software described in the previous section, with 2/3-gram frequencies as the base data and fr+itet as the distance parameters. (For convenience, each score was subtracted from one to convert it back from a dissimilarity into a similarity score.) Next, a matrix of dissimilarities was calculated, using the same parameter settings, between each of the LOB texts and each of the Pentaglossal Corpus texts. This gave a matrix of 500 rows by 113 columns. Then the correlations (Pearson *r* again) between each of the 500 rows of this dissimilarity matrix were

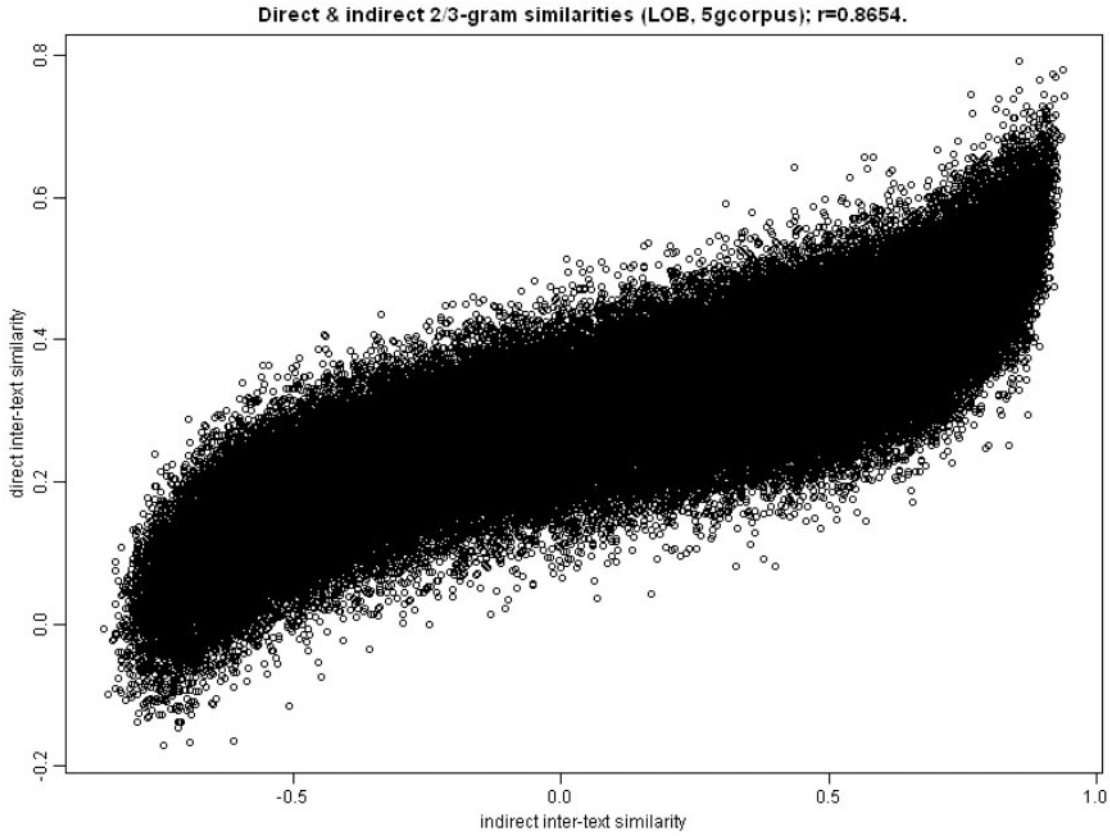

Fig. 4 Comparison of direct and indirect inter-text similarities: LOB versus Pentaglossal Corpus (2/3-grams, fr+itet)

computed, resulting in a 500-by-500 matrix of (indirect) similarities. Finally, the upper triangular sections of both the directly and indirectly calculated similarities were correlated with each other (Pearson *r* yet again). This gave 1,24,750 pairs of similarities in total. The correlation between them was $r = 0.8654$.

Figure 4 is a scatter plot of all these 1,24,750 points, illustrating the relationship between direct and indirect similarities in this case. A correlation of 0.8654 can be interpreted as showing that nearly 75% of the information in the direct similarity matrix is preserved in the indirect similarity matrix, or alternatively that ~25% of the information is lost by resorting to indirect similarity calculation. The diagram shows this relationship as an anti-sigmoid or 'tilted tilde', indicating substantial non-linearity, which could potentially be exploited to predict direct from indirect similarity scores even better.

The same procedure was repeated with the same parameter settings using the Brown Corpus (Kucera and Francis, 1967) as the domain corpus and the LOB corpus as anchor texts. In this case, the resulting correlation of indirectly with directly computed similarities was $r = 0.8308$. These results were taken as sufficiently encouraging to attempt the next step—cross-language similarity estimation.

7 Estimating Inter-Text Similarities Across Languages with Parallel Anchor Texts

To move from monolingual to bilingual similarity-scoring, we make use of the fact that each of the 113 texts in the Pentaglossal Corpus has a

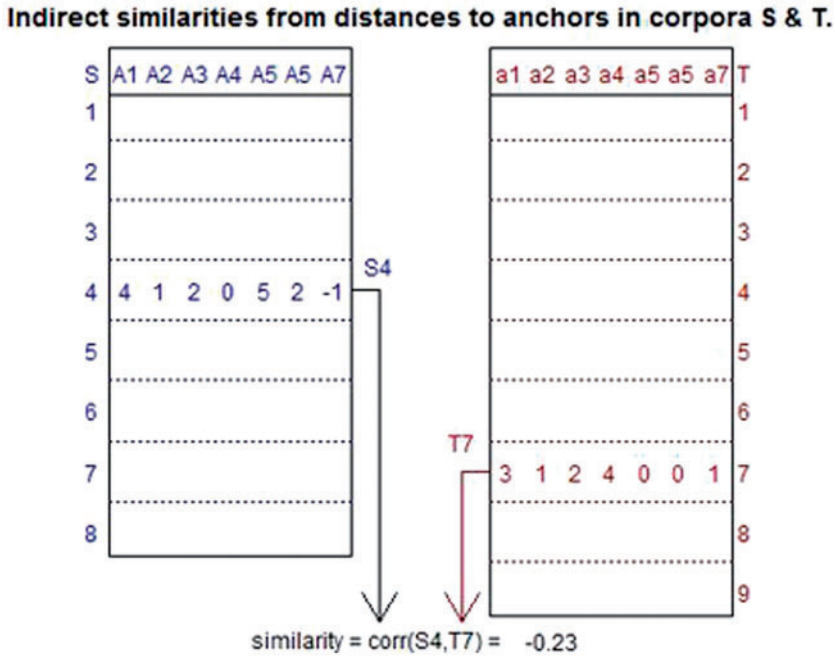

Fig. 5 Cross-language similarity calculation from matrices S and T containing dissimilarity scores to anchor texts

translation equivalent in the other four languages. Thus the dissimilarity profile of a text in English, for example, can be compared with that of a text in German by treating the English and German parallel texts as equivalent anchors.

Figure 5 represents a simple illustration of this concept. Here there are just seven anchor texts in each language. In this example, a1–a7 are translated equivalents in the target language T of anchors A1–A7 in the source language S. Two dissimilarity matrices are shown, S and T, containing dissimilarity scores of eight source-languages texts to anchors A1–A7 and nine target-language texts to anchors a1–a7. For ease of presentation, these are shown as whole numbers. The computed similarity of item 4 in S to item 7 in T is shown.

Of course this is a kind of analogical reasoning that will only give good results to the extent that source-language anchors A1–A7 can be treated as equivalent to target-language anchors a1–a7. To test the procedure empirically, we conducted six cross-language comparisons on the Pentaglossal Corpus: DE-EN, EN-FR, FR-RU, RU-ZH, ZH-DE, and EN-ZH. In each case, a similarity score was computed indirectly (using 1/1-grams with parameter settings: fr+itet) between every text in the source language and each text in the target language. As an outcome measure, these indirect cross-language similarity scores were correlated with the criterial distances established by our human judges, as described in Section 3.

As in this experiment the Pentaglossal Corpus supplied both the domain documents and the anchor texts, a version of the leave-1-out method had to be applied. In fact, this became a 'leave-4-out' method: when comparing texts i and j in languages S and T, the items i and j from anchor set S as well as items i and j from anchor set T had to be disregarded. Thus each indirect similarity calculation used $(n-2) = 111$ anchors.

Table 5 summarizes the results of six cross-language comparisons in which similarities indirectly computed by means of anchor texts were compared with the reference distances of the Pentaglossal Corpus. In this case, more negative correlations are better, as judged distances are being compared with indirectly computed similarities. The median of these cross-language correlations is

−0.8276, which is close to the correlations of indirect with direct monolingual similarities among the Brown, LOB, and Pentaglossal corpora in the previous section (ignoring sign that goes in the right direction in both cases). This suggests that the additional degradation resulting from crossing languages is a relatively minor effect.

Table 5 Cross-language correlations of indirect similarity scores with criterion distances

Language pair	*r*
DE-EN	−0.8334
EN-FR	−0.8179
FR-RU	−0.7669
RU-ZH	−0.8217
ZH-DE	−0.8872
EN-ZH	−0.8623

A visual impression of this relationship is given by Fig. 6, which shows a scatter plot of the German/English comparison data.

8 Discussion

Quantifying similarity between documents is an important subtask in authorship attribution, corpus comparison, information retrieval, and other fields. This study, although limited in scope, has yielded a number of findings relevant to the problem of choosing an effective text similarity-scoring scheme, some of which run counter to established tradition.

8.1 Substantive findings

For estimating monolingual inter-text dissimilarities, it is clear that Euclidean distance, an obvious initial choice when extending case-based reasoning

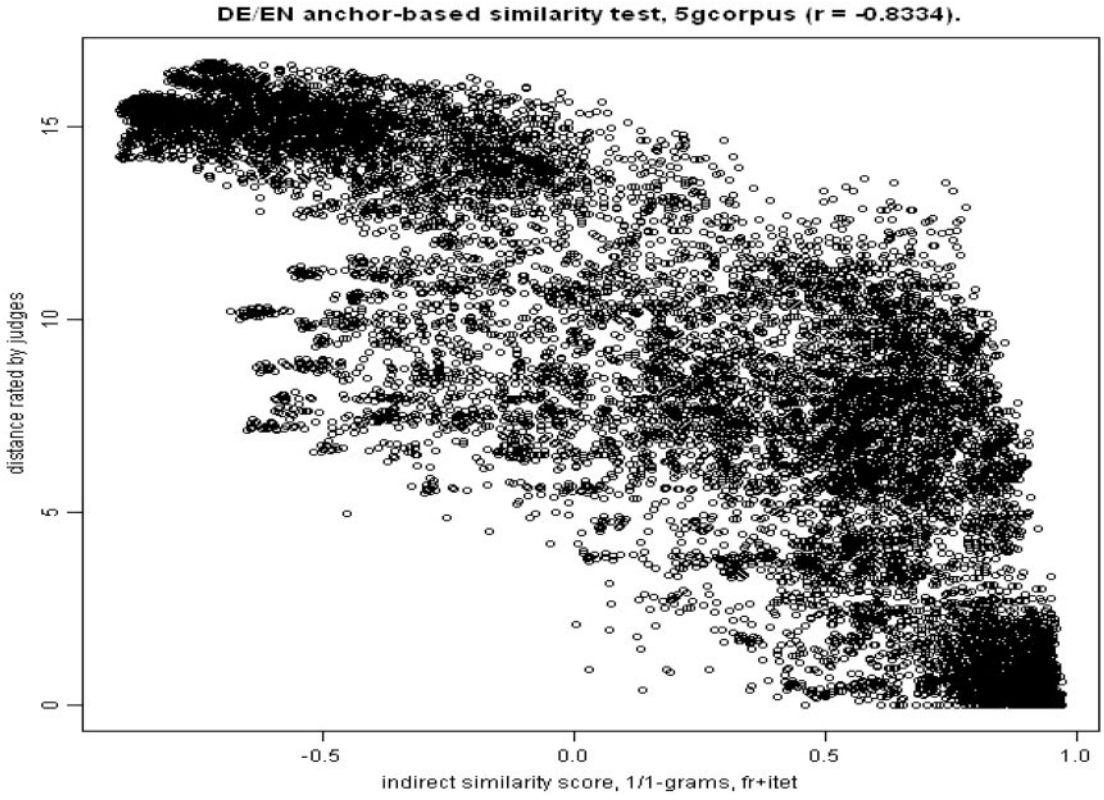

Fig. 6 Cross-language DE-EN similarity/distance test (1/1-grams, fr+itet)

or nearest-neighbour methods to linguistic data, is likely to give poor results. Perhaps more surprising are the results relating to cosine similarity. According to Weiss *et al.* (2005, p. 92): 'Cosine is the default computation for information retrieval and should serve as a benchmark for improvement in any application'. This view on the status of cosine similarity is typical of many discussions of information retrieval and text mining. In the light of this, the fact that cosine similarity is clearly outperformed by alternative correlational measures (particularly Spearman rho and tetrachoric correlation, both of which can be regarded as 'robust' statistics) deserves to be widely broadcast.

Transforming raw frequencies to tf-idf weightings gave good results, but so, somewhat surprisingly, did binarization, especially binarization with reference to median frequency. The best parameter combination involving tf-idf weighting, wt+irho, did not give statistically significantly better results than the best combination that simply used elastigram frequencies, fr+itet. As tf-idf weighting requires examination of an entire reference corpus to compute the inverse document frequencies, whereas binarization merely requires information from the two texts concerned, the latter would seem preferable. It is simpler to compute and unaffected by alterations to any reference corpus.

Disappointingly, the effort to break away from the ubiquitous 'bag-of-words' approach using 2/3-grams, as opposed to 1/1-grams (words), appeared to give no performance advantage, and, certainly, 3/5-grams gave inferior results to either. Only in English did 2/3-grams outperform 1/1-grams, and then only marginally. A plausible explanation for this effect is that longer elastigrams lost more in terms of data sparseness and lack of attribute independence than they gained by tapping into the so-called idiom principle. In any case, the widespread use of words as the standard elements in natural language processing appears to gain empirical support from these findings.

Most notably, this study has demonstrated that similarities between texts in different languages may be calculated to an accuracy comparable with that found in the monolingual case, without bilingual or multilingual lexicons or translation software, using a collection of 'anchor texts' (See Table 5).

8.2 Methodological considerations

In terms of methodology, this study has introduced a viable empirical approach to establishing reference levels for inter-text distance. Many methods of text classification have been proposed and tested over the years, some of which use distance measures to assign documents to categories; however, for the purpose of calibrating such distance measures, we need some way of accommodating the fact that the differences between textual categories are not all equal. The present study has demonstrated a way of doing this, which provides the basis for an intuitively acceptable text-external dissimilarity structure.

We have also shown that inter-text similarity-scoring can be extended from the monolingual case to estimate inter-text similarities across languages by using the concept of 'anchor texts'. Instead of using words (dictionary entries) to build a bridge between languages, we use parallel documents. In a sense, the anchor texts contain an implicit dictionary that we do not have to extract. This means that the methods described in this study do not require pre-existing resources such as lexicons or thesauruses, even though they can be applied to a variety of languages. Nor do they need sophisticated pre-processing such as parsing or tagging. Thus they could easily be applied to under-resourced languages.

Reliable cross-language inter-text similarity-scoring is particularly relevant to the task of building comparable corpora automatically or semi-automatically from documents on the World Wide Web, with a view to improving the quality of statistical machine translation. Statistical machine translation systems are normally trained on parallel corpora, but constructing large parallel corpora is highly resource-intensive. The few that are available to researchers tend to have narrow coverage in terms of domain and genre. Having a robust inter-lingual text similarity-scoring function opens up the prospect of augmenting such narrow training

corpora with a range of near-parallel documents tailored to particular application domains.

Finally, the Pentalossal Corpus itself is a potentially valuable resource available to scholars for studies of the present type and others. In future we hope to improve it in various ways, e.g. by adding other languages, including additional text types, improving some of the less faithful translations, and so on. As it is a public resource, other researchers may also contribute by enhancing it in various ways. Even in its present form, it has demonstrable value. Several huge parallel corpora exist, e.g. Europarl, but they tend to be limited with regard to text variety. A merit of the Pentaglossal Corpus is that it shows that corpora of modest size yet covering a range of registers can still serve a useful purpose. Just how large and how varied such a corpus needs to be to serve as an effective set of anchor texts remains an open question which we hope to address in future.

If the anchor-based approach to cross-language similarity-scoring does become an accepted alternative to the more conventional lexicon-based methodology, it is possible to envisage a development of the Pentaglossal Corpus, or something along similar lines, becoming a standard resource analogous to a multilingual dictionary. Then, users who wish to apply it to a fresh domain could take the generic corpus and add a small number of (parallel) documents from that domain to tune it for a particular application—rather as extra domain-specific terms are added to a standard dictionary to improve the lexicon-based approach to tasks such as machine translation.

Acknowledgements

Several people, as well as the authors, kindly took part in the creation and rating of the Pentaglossal Corpus, including Bogdan Babych, Anne Buckley, Guendalina Gianfranchi, Phoenix Lam, Reinhard Rapp, Fangzhong Su, Edith Ulmer, and James Wilson.

The research reported here has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 248005 (Project TTC).

References

Banchs, R. E. and Costa-Jussà, M. R. (2010). *A Non-Linear Semantic Mapping Technique for Cross-Language Sentence Matching*. Springer: Lecture Notes in Computer Science. *7th International Conference on Natural Language Processing (IceTAL)*, pp. 57–66, Reykjavik, Iceland, August 2010.

Biber, D., Conrad, S., and Cortes, V. (2004). If you look at: lexical bundles in university teaching and textbooks. *Applied Linguistics*, **25**(3): 371–405.

Bross, I. (1958). How to use ridit analysis. *Biometrics*, **14**: 18–38.

Chen, J. and Bau, Y. (2009). Cross-language search: the case of Google language tools. *First Monday*, **14**(3): 2009.

Cheng, W., Greaves, C., and Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, **11**(4): 411–33.

Chowdhury, A., Frieder, O., Grossman, D., and McCabe, M. (2002). Collection statistics for fast duplicate document detection. *ACM Transactions on Information Systems*, **20**(2): 171–91.

Clough, P. and Gaizauskas, R. (2009). Corpora and Text Re-use. In Ludeling, A., Kyto, M., and McEnery, A. (eds), *Handbook of Corpus Linguistics*. Berlin: Mouton de Gruyter, pp. 1249–71.

Everitt, B. (ed.) (1998). *The Cambridge Dictionary of Statistics*. Cambridge: CUP.

Freelon, D.G. (2010). ReCal: inter-coder reliability as a web service. *International Journal of Internet Science*, **5**(1): 20–33.

Hofland, K. and Johansson, S. (1982). *Word frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities/London: Longman.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics*, **15**(3): 651–74.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, **28**(1): 11–21.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, **1**(3): 233–334.

Kapur, J. and Kesavan, H. (1992). *Entropy Optimization Principles with Applications*. Boston: Academic Press.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, **6**(1): 1–37.

Koehn, P. (2005). Europarl: a parallel corpus for statistical machine translation. *Proceedings of the MT Summit*. Phuket, Thailand, September 2005.

Kripendorff, K. (2004). *Content Analysis: An Introduction to its Methodology*. Thousand Oaks, CA: Sage.

Kucera, H. and Francis, W. (1967). *Computational Analysis of Present-day American English*. Providence, RI: Brown University Press.

Li, B. and Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. *Proceeding of the 23rd International Conference on Computational Linguistics* (*Coling 2010*). Beijing, China, August 2010.

Min, H. and McCarthy, P. (2010). Identifying varietals in the discourse of American and Korean scientists. *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference* (*FLAIRS 2010*). Daytona, FL.

Potthast, M., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011). Cross-language plagiarism detection. *Language Resources and Evaluation*, **45**(1): 45–62.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rajman, M. and Hartley, T. (2001). Automatically predicting MT systems rankings compatible with fluency, adequacy or informativeness scores. *Proceedings of the 4th ISLE Workshop on MT Evaluation*. Santiago de Compostela, pp. 29–34, September 2001.

Rosengren, I. (1971). The quantitative concept of language and its relation to the structure of frequency dictionaries. *Etudes de Linguistique Appliquée*, **1**: 103–27.

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Sammon, J. (1969). A non-linear mapping for data structure analysis. *IEEE Transactions on Computers*, **18**: 401–9.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: OUP.

Steinberger, R., Pouliquen, B., and Hagman, J. (2002). Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. *Computational Linguistics and Intelligent Text Processing, Third International Conference*. Mexico City, Mexico, pp. 101–21, February 2002.

Su, F. and Babych, B. (2012). Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-)parallel translation equivalents. *Proceedings of the 13th Conference of EACL*, pp. 10–19, Avignon, France, April 2012.

Upton, G. and Cook, I. (2008). *The Oxford Dictionary of Statistics*. Oxford: OUP.

Weiss, S., Indurkhya, N., Zhang, T., and Damerau, F. (2005). *Text Mining*. New York: Springer.

Wu, Z., Markert, K., and Sharoff, S. (2010). Fine-grained genre classification using structural learning algorithms. *Proceedings of the ACL 2010*, pp. 749–59. Uppsala, July 2010.

Appendix

Attribute code	Question to be answered
A1. Polemic	To what extent does the text seek to persuade the reader to support (or renounce) an opinion or point of view?
A2. Corp	To what extent is corporate authorial responsibility indicated? (Mainly/Wholly for texts clearly produced on behalf of an organization without named authors. None if a named individual or named individuals indicate authorship.)
A3. Emotive	To what extent is the text concerned with expressing feelings or emotions? (None for neutral explanations, descriptions and/or reportage.)
A4. Fictive	To what extent is the text's content fictional? (None if you judge it to be factual/informative.)
A5. Flippant	To what extent is the text light-hearted, i.e. aimed mainly at amusing or entertaining the reader? (None if it appears earnest or serious.)
A6. Informal	To what extent is the text's content written in an informal style, using colloquialism and/or slang (as opposed to the 'standard' or 'prestige' variety of language)?
A7. Tutorial	To what extent does the aim of the text seem to be to teach the reader how to do something (e.g. a tutorial)?
A8. News	To what extent does the text appear to be a news story such as might be found in a newsletter, newspaper, magazine, or other periodical, i.e. a report of recent events (recent at the time of writing at any rate)?
A9. Legalist	To what extent does the text lay down a contract or specify a set of regulations? (Includes copyright/copyleft notices.)
A10. Locutive	To what extent does the text represent spoken discourse (including fictional "dialogues" and monologues, as well as scripts written to be spoken)?
A11. Personal	To what extent is the text is written from a first-person point of view?
A12. Compuff	To what extent does the text promote a commercial product or service?
A13. Ideopuff	To what extent is the text intended to promote a political movement, party, religious faith, or other non-commercial cause (i.e. any promotion of not-for-profit causes)?
A14. Scitech	To what extent would you categorize the text's subject-matter as belonging in the field of Science, Technology, and/or Engineering (as opposed to the Arts, Humanities, and/or Social Studies)?
A15. Specialist	To what extent does the text, in your opinion, require background knowledge of a specialized subject area (such as would not be expected of the so-called 'general reader') in order to be comprehensible?
A16. Oral	To what extent do you believe that the text originates from spoken discourse?
A17. Modern	To what extent do you judge the text to be modern?

Rating Levels:
0 none or hardly at all;
0.5 slightly;
1 somewhat or partly;
2 strongly or very much so.